A Single-letter Characterization of Optimal Noisy Compressed Sensing

Dongning Guo
Electrical Engineering & Computer Science Dept.
Northwestern University
Evanston, IL 60208, USA

Dror Baron and Shlomo Shamai (Shitz)
Electrical Engineering Dept.
Technion - Israel Institute of Technology
Haifa 32000, Israel

***Abstract*—Compressed sensing deals with the reconstruction of a high-dimensional signal from far fewer linear measurements, where the signal is known to admit a sparse representation in a certain linear space. The asymptotic scaling of the number of measurements needed for reconstruction as the dimension of the signal increases has been studied extensively. This work takes a fundamental perspective on the problem of inferring about individual elements of the sparse signal given the measurements, where the dimensions of the system become increasingly large. Using the replica method, the outcome of inferring about any fixed collection of signal elements is shown to be asymptotically decoupled, i.e., those elements become independent conditioned on the measurements. Furthermore, the problem of inferring about each signal element admits a *single-letter* characterization in the sense that the posterior distribution of the element, which is a sufficient statistic, becomes asymptotically identical to the posterior of inferring about the same element in scalar Gaussian noise. The result leads to simple characterization of all other elemental metrics of the compressed sensing problem, such as the mean squared error and the error probability for reconstructing the support set of the sparse signal. Finally, the single-letter characterization is rigorously justified in the special case of sparse measurement matrices where belief propagation becomes asymptotically optimal.**

I. Introduction

The representation and reconstruction of sparse high-dimensional signals from far fewer linear measurements has received much attention in recent years. Donoho [1] and Candés and Tao [2] showed that compressed sensing (CS) based on convex programming is asymptotically optimal in the sense that reconstruction can be achieved using essentially as few measurements as any other estimator would need. A large body of literature has since emerged to address the theoretical limits as well as practical issues in CS. Most analytical treatises on CS study the problem of how many noiseless or noisy measurements are asymptotically sufficient for reconstruction of the sparse (input) signal under some given scalar metric, e.g., the mean squared error or the probability or amount of errors for reconstructing the support set of the sparse signal.

The goal of this paper is to address the following fundamental question pertaining to noisy CS: What can one infer about an individual element of the sparse signal based on the measurements? To make progress, a Bayesian framework for statistical inference with noisy measurements is considered, where the input statistics are assumed known to the estimator. For each input element, it is clear that the posterior distribution of the element conditioned on the measurements is a sufficient statistic, so that the problem boils down to characterizing this posterior. In fact, the posterior is itself random, as it is a function of the measurements and the measurement matrix. Evidently, for a system of given size, the posterior of an individual element has a complicated structure in general.

In this paper, we put forth a simple *single-letter characterization* of the posterior of each individual element conditioned on the measurements in a certain large-system limit. It is shown that the asymptotic *distribution of the (random) posterior* is surprisingly simple. In fact, as the input dimensionality and the number of measurements both increase, *this posterior becomes statistically identical to the posterior of a scalar Gaussian channel*[1] whose signal-to-noise ratio (SNR) can be obtained by solving a fixed-point equation. Taking another perspective, we can say that whatever one can infer about an input element of the sparse signal based on the measurements is asymptotically identical to what one can infer about the same element if all other input elements were zero, but the measurements

The work of D. Guo has been supported by the NSF under grant CCF-0644344. The work of D. Baron and S. Shamai has been supported by the Israel Science Foundation and by the Technion Fund for Research.

[1]Hence the term single-letter characterization.

were noisier. Another contribution of this work is to show that the equivalent scalar Gaussian channels for the input elements can be essentially *decoupled*. That is, conditioned on the measurements, any fixed set of input elements are asymptotically independent in the large-system limit.

The single-letter characterization of the marginal posterior distribution leads to a simple characterization of all other elemental metrics of the CS problem, such as the minimum mean-square error (MMSE), the error probability, the entropy, etc. This result is convenient for many practical purposes, for example, to determine the number of measurements and the SNR required for achieving a certain quality of reconstruction. We note that the results in this paper also advance the understanding of the fundamental nature of noisy CS by describing a boundary between what is physically possible and what is not. Another sharp characterization of phase transition deals only with noiseless measurements [3], [4]. The result in this paper is thus sharper than many other results on noisy CS obtained using the restricted isometry property developed in [5]. There have been several other works on the information-theoretic performance bounds of CS, e.g., [6]–[11]. The unique asymptotic regime considered in this work allows an *exact* characterization of the performance of noisy CS. In fact the results here often offer a better approximation for the performance of finite-size systems than the existing bounds.

The techniques used to develop the results in this paper are generally applicable to large linear systems, including code-division multiple access (CDMA) systems, multiple-antenna channels, as well as CS systems, where the distinct feature of the latter is the sparsity of the input. In particular, the instrumental replica method was invented to analyze macroscopic properties of spin glasses in statistical mechanics [12]. Tanaka [13] first used the replica mothod to analyze the optimal error probability of CDMA with binary inputs. Guo and Verdú [14], [15] generalized Tanaka's result to *arbitrary* inputs, which is applicable to sparse inputs (see also [16]).[2] A simple performance characterization of the large linear system using a bank of scalar channels was also developed in [15] for the first time. Indeed one of the goals of this paper is to adapt the main findings of [15] to the CS application, so that the results can be easily applied in the new context. Furthermore, extensions to the previous results are presented, which include: 1) a formal statement of the asymptotic decoupling of the posterior of the inputs, and 2) a connection between the optimal performance and the performance achieved by iterative belief propagation. This latter connection has previously been established in the CDMA context [18].

We note that a recent independent work by Rangan, Fletcher and Goyal [10] also applies the results and techniques of Guo and Verdú [15] to develop a similar characterization of the performance of the maximum *a posteriori* (MAP) estimator and several fast compressed sensing algorithms, such as basis pursuit. In particular, the MAP estimator is treated as the limit of a sequence of conditional mean estimators studied in [15],[3] which is are referred to as MMSE estimators in [10], so that the results in [15] can be used to obtain the limiting performance of the MAP estimator. This paper and [10] are thus related, although the focus of [10] is the MAP estimator and several suboptimal estimators, whereas this paper puts the emphasis on the posterior distribution of the input elements, which is a sufficient statistic. Both works, however, advocate the simple characterization obtained via the replica method.

A final contribution of this paper is that we draw the link between optimal detection and belief propagation (BP) in the context of CS. The single-letter characterization is rigorously justified in the special case of sparse measurement matrices. It is found that sparse measurement matrices perform just as well; and BP is asymptotically optimal in case of sparse mesurement matrix.

The remainder of the paper is organized as follows. Section II describes the system model. A set of results for the case of dense measurement matrices is presented in Section III. The counterpart for the case of sparse measurement matrices is shown in Section IV. In Section V, we discuss the performance for a special type of inputs. Section VI concludes the paper.

II. System Model

Consider a (stochastically) sparse signal $\boldsymbol{X}$ in a known N-dimensional space in the sense that *a priori* most of the entries of its vector representation are zero. Specifically, for each $n = 1, \ldots, N$, let the n-th entry of $\boldsymbol{X}$ be $X_n = B_n U_n$, where B_n is Bernoulli with probability ϵ to be 1, and $U_n \sim P_U$, an arbitrary distribution with $\mathsf{E}\left\{U^2\right\} = 1$ and arbitrary expected value. The distribution of X_n is thus a mixture of P_U and a point

[2] In a contemporary work [17], Müller used the replica method technique to evaluate the performance of systems with many antennas.

[3] Such "hardening" techniques for achieving the MAP estimator have been used by Tanaka [13] and also in [15] to deal with the maximum-likelihood detector and the decorrelator.

mass at 0; we call X_n where $B_n = 1$ an active element. Moreover, it is assumed that $B_1, \ldots, B_N, U_1, \ldots, U_N$ are mutually independent.

Suppose that M random linear measurements are taken, where the m-th measurement Y_m can be regarded as an inner product of the signal and a measurement vector $[S_{m1}, S_{m2}, \ldots, S_{mN}]$. It is assumed that the measurements are contaminated by additive noise, so that Y_m can be expressed as

$$Y_m = \sqrt{\frac{\gamma}{M}} \sum_{n=1}^{N} S_{mn} X_n + W_m \tag{1}$$

where $W_m \sim \mathcal{N}(0,1)$ are independent and identically distributed (i.i.d.) standard Gaussian for $m = 1, \ldots, M$. It is further assumed that the measurement vectors are generated randomly, so that the weights S_{mn} can be regarded as i.i.d. with distribution P_S, which is of zero mean and unit variance. It is easy to see that the average SNR of each measurement is γ.

The statistical system model is completely described by $(N, M, \epsilon, \gamma, P_U, P_S)$, i.e., the dimension of the signal, the number of measurements, the sparsity, the SNR, and the distributions of the nonzero inputs and measurement coefficients. The performance of such a system can of course be evaluated for arbitrary parameters, but the result is often too complex to provide any insight. In order to make progress, this paper considers the following *large-system limit*: Fix $(\epsilon, \gamma, P_U, P_S)$ but let $N, M \to \infty$ with

$$\frac{M}{\epsilon N} \to \mu \tag{2}$$

where μ is a positive constant. Clearly, ϵN is the average number of active input elements so that μ denotes the limit of the average number of measurements per active element.

For brevity, the vector of measurements described by (1) can be expressed as

$$\boldsymbol{Y} = \sqrt{\gamma}\, \underline{\boldsymbol{S}} \boldsymbol{X} + \boldsymbol{W} \tag{3}$$

where $\boldsymbol{W}$ consists of independent standard Gaussian entries, and with slight abuse of notation the (m, n) entry of $\underline{\boldsymbol{S}}$ is set to $S_{mn}/\sqrt{M}$. In light of (2), it is easy to see that each column of $\underline{\boldsymbol{S}}$ has unit energy in the large-system limit. We note in passing that the noisy linear measurement model appears in numerous other problems, including CDMA [15], multiple-input multiple-output (MIMO) systems [17], and machine learning [19].

The large-system limit evaluated in this paper obviously differs from many other CS works, where the number of measurements is often on the order of the logarithm of the signal dimensionality, e.g., $M = O(K \log N)$ where K is the cardinality of the support set of the sparse signal (cf., [1], [7]). Having a fixed ratio between the length of the signal (N) and the number of measurements (M) is mathematically convenient. Studying this asymptotic regime provides equally abundant insights as provided by other regimes, since after all, the goal is to provide a good approximation to systems of finite dimension (N, M) in practice.

III. Dense Measurement Matrix

The success of compressed sensing (CS) relies on reliable reconstruction of the original sparse signal $\boldsymbol{X}$ despite the noisy measurements. Typical results in the existing CS literature address the ℓ_2 norm of the estimation error for $\boldsymbol{X}$ [6], [7]. Some other results provide scaling laws—typically what should the number of measurements be in terms of the order of the length of $\boldsymbol{X}$ for reliable reconstruction. In this work we provide an accurate characterization of the performance in terms of estimating each individual element of the vector $\boldsymbol{X}$.

The model (3) has been studied in the large-system limit [15] in the context of CDMA. Indeed the CS model takes the same mathematical form as that of CDMA, except that the prior distribution of the input elements puts a large probability mass at 0. The general result in [15] is applicable to such cases in the large-system limit defined in Section II. In the following, we first give an example of the consequence of the general result in the special case where the sparse input vector consists of i.i.d. Bernoulli random variables. We then describe the main result of this paper for a dense measurement matrix in full generality and discuss its implications.

A. A Special Case: Bernoulli Inputs

The special case of Bernoulli inputs has been studied in the CS literature, (cf., [8]). Consider the case where the input elements $X_n = B_n$ are Bernoulli with parameter ϵ. Detection of each element B_n is thus a hypothesis testing problem with two hypotheses corresponding to $B_n = 0$ and $B_n = 1$, respectively:

$$H_0: \quad \boldsymbol{Y} = \sqrt{\gamma} \sum_{i=1, i\neq n}^{N} \boldsymbol{S}_i B_i + \boldsymbol{W} \tag{4}$$

$$H_1: \quad \boldsymbol{Y} = \sqrt{\gamma} \sum_{i=1, i\neq n}^{N} \boldsymbol{S}_i B_i + \boldsymbol{W} + \sqrt{\gamma}\, \boldsymbol{S}_1 \tag{5}$$

where $\boldsymbol{S}_n$ is the n-th column of the measurement matrix. There are two types of errors: misses and false alarms. The following result characterizes the

fundamental trade-off between the probabilities of these two types of errors for recovering each input element. We emphasize that our *claims* are based on heuristic yet well-accepted arguments (such as the replica method [12], [13], [16]) from statistical mechanics. Additionally, we note that Claim 1 follows directly from Claim 2.

Claim 1: Let $\hat{B}_n$ denote the reconstruction of the element B_n. In the large-system limit ($N, M \to \infty$ with $M/N \to \epsilon\mu$), the optimal trade-off between the probabilities of the two types of errors are described by the following formulas parameterized by $t \in \mathbb{R}$:

$$\mathsf{P}\{\hat{B}_n = 1 | B_n = 0\} = Q\left(t\right) \triangleq \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} \, \mathrm{d}u \tag{6}$$

$$\mathsf{P}\{\hat{B}_n = 0 | B_n = 1\} = Q\left(\sqrt{\eta\gamma} - t\right) \tag{7}$$

where η is some constant in $(0, 1)$, which depends on (ϵ, μ, γ) but not on t.

The trade-off described by (6) and (7) is surprisingly simple. In fact, it is identical to the trade-off associated with the following hypothesis testing problem:

$$H_1: \quad Y = \sqrt{\eta\gamma} + W \tag{8}$$

$$H_0: \quad Y = W \tag{9}$$

where $W \sim \mathcal{N}(0, 1)$ is standard Gaussian. That is to say, the optimal performance of reconstructing the element B_n based on a large number of measurements is no different from that of recovering B_n based on a scalar measurement contaminated by additive Gaussian noise. The parameter $\eta \in (0, 1)$ acts as a degradation of the SNR, and we refer to it as the *energy efficiency* or simply the *efficiency* of the CS system. The efficiency is determined in Section III-B. Note that the probabilty of false alarms and the probability of misses are decreasing and increasing functions of t, respectively. Moreover, if one of the probabilities is driven to zero then the other probability necessarily approaches one.

The role of the efficiency η is also quite simple. Let $n \in \{1, \ldots, N\}$ be fixed. For a moment consider a model also expressed by (3) with the same statistics except that all but the n-th entry of the input is suppressed *a priori*, i.e., $X_{n'} = 0$ for all $n' \neq n$. Let $X_n = B_n$ still be Bernoulli with prameter ϵ. A sufficient statistic of $(\boldsymbol{Y}, \underline{\boldsymbol{S}})$ for B_n is obtained by matched filtering with respect to the n-th column of $\underline{\boldsymbol{S}}$, which can be expressed in the large-system limit as $Z_n = \sqrt{\gamma} B_n + W$ with $W \sim \mathcal{N}(0, 1)$ because each column of $\underline{\boldsymbol{S}}$ has unit energy asymptotically. Thus detection of B_n can be regarded as a hypothesis testing problem described by (8) and (9) with $\eta = 1$. Therefore, detecting B_n via CS is analogous to detecting this input element with all other elements suppressed, but based on a noisier observation (where the SNR is degraded by a factor of η).

Finally, we note that Claim 1 offers a more precise performance characterization than the information-theoretic bounds developed by Aeron et al. [11], which consider the same large-system limits.

B. Single-letter Characterization: General Inputs

For general inputs, the problem of reconstructing each input element is more involved than the testing of two hypotheses. We note that a sufficient statistic of $(\boldsymbol{Y}, \underline{\boldsymbol{S}})$ for the input element X_n is the posterior distribution $P_{X_n|\boldsymbol{Y},\underline{\boldsymbol{S}}}(\cdot|\boldsymbol{Y}, \underline{\boldsymbol{S}})$. If we were able to describe this posterior exactly, then everything would be known about the quality of any kind of inference one wishes to make about X_n. This is of course in general an infeasible task because of the complicated structure of the posterior, which is a function of $\boldsymbol{Y}$ and $\underline{\boldsymbol{S}}$. Surprisingly, it turns out that the posterior admits a simple characterization in the large-system limit, which is described as a consequence of Claim 2 in [15] below.

An important role is played here by the MMSE of estimating a signal through a Gaussian channel. Specifically, we denote the MMSE for estimating an arbitrary real-valued random variable X based on the value of $\sqrt{\gamma}\, X + W$ by

$$\mathsf{mmse}(P_X, \gamma) = \mathsf{E}\left\{\left(X - \mathsf{E}\left\{X | \sqrt{\gamma}\, X + W\right\}\right)^2\right\} \tag{10}$$

where $W \sim \mathcal{N}(0, 1)$ is standard Gaussian, and γ represents the SNR gain of the channel. Evidently, $\mathsf{mmse}(P_X, \gamma)$ is equal to the variance of X at $\gamma = 0$ and vanishes monotonically as $\gamma \to \infty$.

The following result is a single-letter characterization of the compressed sensing problem modeled by (3) in the large-system limit.

Claim 2: As far as inferring about X_n is concerned, in the large-system limit ($N, M \to \infty$ with $M/N \to \epsilon\mu$), the observation of $(\boldsymbol{Y}, \underline{\boldsymbol{S}})$ becomes statistically equivalent to observing X_n with additive Gaussian noise of variance $(\eta\gamma)^{-1}$ for some $\eta \in (0, 1)$, or equivalently, observing some $Z_n \sim \mathcal{N}(\sqrt{\eta\gamma}\, X_n, 1)$. That is, conditioned on the actual value of $X_n = x$, the posterior distribution converges in distribution as the system becomes large:

$$P_{X_n|\boldsymbol{Y},\underline{\boldsymbol{S}}}\big(\,\cdot\,\big|\boldsymbol{Y}, \underline{\boldsymbol{S}}\big) \xrightarrow{D} P_{X_n|Z_n}(\cdot|Z_n). \tag{11}$$

The parameter η satisfies the following fixed-point equation

$$\eta^{-1} = 1 + \frac{\gamma}{\epsilon\mu}\, \mathsf{mmse}(P_X, \eta\gamma). \tag{12}$$

In case of multiple solutions to (12), η is chosen as the one that minimizes

$$I(X; \sqrt{\eta\gamma}\, X + N) + \frac{\epsilon\mu}{2}(\eta - 1 - \log \eta).$$

We first note that $P_{X_n|Z_n}(\cdot|Z_n)$ is a random probability measure on $\mathbb{R}$, which depends on Z_n. Formula (11) states that the sequence of random probability measures $P_{X_n|\boldsymbol{Y},\underline{\boldsymbol{S}}}(\cdot|\boldsymbol{Y},\underline{\boldsymbol{S}})$ converges to the probability measure $P_{X_n|Z_n}(\cdot|Z_n)$ in distribution. For concreteness, (11) is equivalent to convergence of the cumulative distribution functions (cdfs) in distribution, i.e.,

$$\begin{aligned} &\mathsf{P}\{X_n \le x \big| \boldsymbol{Y}, \underline{\boldsymbol{S}}\} \\ &\xrightarrow{D} \mathsf{P}\{X_n \le x | Z_n\} \\ &= \frac{\int_{-\infty}^{x} \exp\left[-\frac{1}{2}(Z_n - \sqrt{\eta\gamma}\, u)^2\right] \mathrm{d}P_X(u)}{\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(Z_n - \sqrt{\eta\gamma} u)^2\right] \mathrm{d}P_X(u)} \end{aligned} \tag{13}$$

for every x where the cdf $P_X(\cdot)$ is continuous. Note that here the conditional distribution functions $P_{X_n|Z_n}(x|z)$ are identical for all $n = 1, \ldots, N$. For notational convenience, let $P_{X_n|Z_n}$ be denoted by $P_{X|Z}$ from this point on.

The essence of the general result given by Claim 2 is to characterize the posterior for each input element by the simple posterior of a scalar Gaussian channel. This principle is a special case of the general result originally developed in [15], [16]. Thus there is no need for a separate proof. An illustration of the result is shown in Figure 1. A simple consequence of Claim 2 is the following result on the elemental estimation error.

Corollary 1: The MMSE of estimating each input element is $\mathsf{mmse}(P_X, \eta\gamma)$. The MMSE of estimating the input vector $\boldsymbol{X}$ is $\mathsf{mmse}(P_X, \eta\gamma)$ per dimension.

From Claim 2, it is easy to see that, in the special case where $X_n = B_n$, the problem of inferring about X_n via CS is as characterized in Section III-A. Moreover, the efficiency η can be determined from the fixed-point equation (12), where the MMSE in this case admits the following expression:

$$\begin{aligned} &\mathsf{mmse}(\epsilon\delta(1) + (1-\epsilon)\delta(0), \gamma) \\ &= \epsilon - \frac{\epsilon^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{2y\sqrt{\gamma} - \gamma}}{\epsilon\, e^{y\sqrt{\gamma} - \frac{\gamma}{2}} + 1 - \epsilon} e^{-\frac{y^2}{2}} \mathrm{d}y \ . \end{aligned} \tag{14}$$

C. Decoupling of Input Elements

Oftentimes one is interested in inferring about all or a subset of input elements. The question becomes what is the joint posterior distribution of the input elements given the measurements. The joint posterior is in general very complex, but in the large-system limit, the following decoupling result can be shown using the replica method. The detailed proof is omitted here due to space limitations.

Claim 3: Let the efficiency η be the same as determined by Claim 2. Consider an arbitrary but fixed number L of input elements $(X_{n_1}, \ldots, X_{n_L})$. As $N, M \to \infty$ with $M/N \to \epsilon\mu$,

$$\begin{aligned} &P_{X_{n_1},\ldots,X_{n_L}|\boldsymbol{Y},\underline{\boldsymbol{S}}}\big(x_1, \ldots, x_L \big| \boldsymbol{Y}, \underline{\boldsymbol{S}}\big) \\ &\qquad \xrightarrow{D} \prod_{i=1}^{L} P_{X|Z}(x_i|Z_{n_i}) \end{aligned} \tag{15}$$

for all $x_1, \ldots, x_L$, where $Z_{n_i} = \sqrt{\eta\gamma}\, X_{n_i} + W_i$ with i.i.d. $W_i \sim \mathcal{N}(0,1)$ and $P_{X|Z}$ is defined as in Claim 2.

We caution that the above asymptotic decoupling concerns a constant number L of input elements and cannot be extended to the joint posterior of all input elements (their population $N \to \infty$). The decoupling of the posterior suggests that the decision made on one input element is asymptotically uncorrelated with that on other elements, and so Claim 2 is a special case of Claim 3. Finally, an illustration of decoupling for the special case of sparse measurement matrices appears in Figure 2.

IV. Sparse Measurement Matrix

Consider a scenario where the measurement matrix is sparse so that each input element affects only a small fraction of the measurements. The sparsity of the measurement may be due to the nature of the system or by design, for example to CS fast computation [20]. It turns out that, under many circumstances, the preceding decoupling results obtained using the replica method can be proved rigorously. Moreover, in those cases, belief propagation (BP) detection can be shown to be asymptotically optimal in the sense that it can compute the marginal posterior probability of each input element.

For concreteness, we consider a sequence of ensembles of measurement matrices indexed by the input dimensionality N. Let the number of measurements M be a function of N such that $M/N \to \epsilon\mu$ as $N \to \infty$ (as in Section II). Let the matrix sparsity $q \in (0,1)$ satisfy $qM \to \infty$ and $qM^a \to 0$ for all $a < 1$ (for example, the sparsity can be $q = (\log M)/M$). For each (N, M), a matrix $\underline{\boldsymbol{S}}$ randomly drawn from the ensemble would have the following statistics: All of its entries are i.i.d. with probability $1-q$ to be set to 0, and otherwise follow the distribution P_S, which is of zero mean, unit variance, and finite fourth-order moment. The measurement matrix is $\underline{\boldsymbol{S}} = \frac{1}{\sqrt{qM}}(S_{mn})$, and so every column has unit energy on average. Clearly, each input element

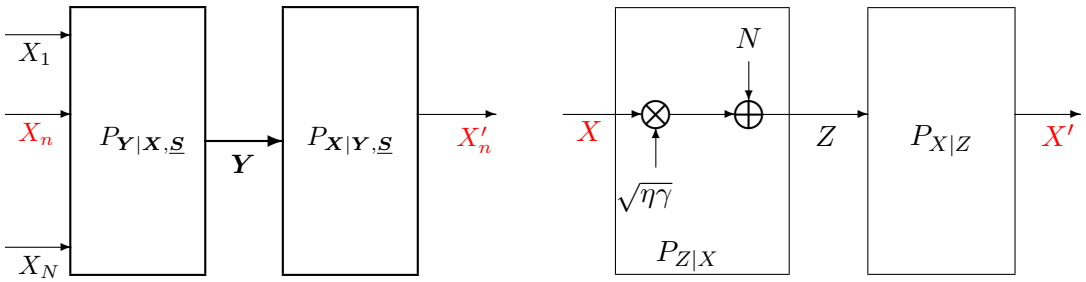

Fig. 1. For large systems, the posterior of any input X_n given the measurements $\boldsymbol{Y} = \sqrt{\gamma}\,\underline{\boldsymbol{S}}\boldsymbol{X} + \boldsymbol{W}$ and the measurement matrix $\underline{\boldsymbol{S}}$ is statistically identical to the posterior of X given $Z = \sqrt{\eta\gamma}X + N$, where η satisfies (12). In the graph, if the inputs X and X_n are identical, then the statistics of X' and X'_n are indistinguishable.

affects qM measurements on average, which becomes large but increasingly sparse as $M \to \infty$. Under certain circumstances, Claims 2 and 3 can be rigorously shown. Hence the following result, whose proof is omitted due to space limitations.

Theorem 1: Consider the system described by (3) in the large-sparse-system limit, i.e., where $N, M \to \infty$ with $M/N \to \epsilon\mu$, $qM \to \infty$ and $qM^a \to 0$, $\forall a < 1$. Suppose the fixed-point equation (12) for the efficiency η has a unique solution, then for fixed $(n_1, \ldots, n_L)$,

$$P_{X_{n_1},\ldots,X_{n_L}|\boldsymbol{Y},\underline{\boldsymbol{S}}}\big(x_1,\ldots,x_L\big|\boldsymbol{Y},\underline{\boldsymbol{S}}\big) \xrightarrow{D} \prod_{i=1}^{L} P_{X|Z}(x_i|Z_{n_i}) \tag{16}$$

for all $x_1, \ldots, x_L$, where $Z_{n_i} = \sqrt{\eta\gamma}\,x_{n_i} + W_i$ with i.i.d. $W_i \sim \mathcal{N}(0,1)$ and $P_{X|Z}$ is defined as in Claim 2.

We would like to point out that the decoupling result is related to previous decoupling results for non-sparse inputs due to Guo and Wang [18] and Montanari [21].

Note that the system (3) can be described using a factor graph. In particular, the joint distribution of all input elements and measurements can be factorized to a product of one conditional distribution term for each measurement and one prior distribution term for each input element. The sparsity of the measurement matrix is such that the factor graph is locally tree-like, in the sense that as the system size becomes large, the probability of having cycles shorter than any given length vanishes. It is well known that BP computes the exact marginal posterior probability distribution for cycle-free graphs. After t iterations, the output of BP is a posterior distribution for X_k computed based on all measurements within distance $2t-1$ of X_k on the factor graph, denoted by $\boldsymbol{Y}_k^{(t)}$ [18]. With slight abuse of notation, let $P_{X_k}^{\text{bp}}(\cdot|\boldsymbol{Y}_k^{(t)},\underline{\boldsymbol{S}})$ denote the output cdf of BP, which is the approximate posterior of X_k given $\boldsymbol{Y}_k^{(t)}$ and the measurement matrix $\underline{\boldsymbol{S}}$.

Theorem 2: Consider the same model and conditions as in Theorem 1. For any number of iterations t, the posterior obtained by BP converges:

$$P_{X_k}^{\text{bp}}\big(x\big|\boldsymbol{Y}_k^{(t)},\underline{\boldsymbol{S}}\big) \to P_{X|Z}\big(x\big|h\big(\boldsymbol{Y}_k^{(t)},\underline{\boldsymbol{S}}\big)\big)$$

in probability in the large-sparse-system limit for some function $h(\cdot)$ such that $h\big(\boldsymbol{Y}_k^{(t)},\underline{\boldsymbol{S}}\big) \sim \mathcal{N}\big(\sqrt{\eta^{(t)}\gamma}\,x, 1\big)$ conditioned on $X_k = x$, where $\eta^{(t)}$ is the result of the following iterative formula

$$\left(\eta^{(t+1)}\right)^{-1} = 1 + \frac{\gamma}{\epsilon\mu}\,\mathsf{mmse}\left(P_X, \eta^{(t)}\gamma\right). \tag{17}$$

with $\eta^{(0)} = 0$.

It is clear that the fixed point of (17) satisfies (12). In light of Theorem 2, the posterior distribution in large sparse systems can essentially be obtained by BP in the special case where the solution to (12) is unique. It turns out that the equation may have up to three fixed points [13], [16]. In such cases, the optimal performance achieved by any estimator is not known to admit a single-letter characterization, but the performance can be shown to be sandwiched between the single-letter characterization with the smallest and the largest of the fixed points.

To illustrate the decoupling effect given by Theorem 1, we ran a CS reconstruction algorithm based on BP [20]. As shown in Theorem 2, combining BP with a sparse measurement matrix offers asymptotically optimal estimation in addition to computational efficiency [20]. Figure 2 illustrates that the posterior for X_1 given the observations $(\boldsymbol{Y},\underline{\boldsymbol{S}})$ is invariant of the posterior for X_2. The numerical results also illustrate the decoupling effect described in Claim 3, although it would be infeasible to simulate exact a posterior estimation, owing to the prohibitive complexity.

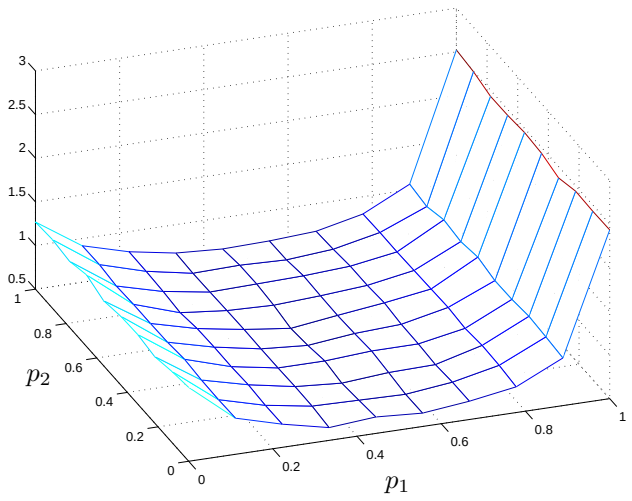

Fig. 2. The graph shows the normalized histogram of $p_1 = \mathsf{P}\{X_1 = 1|\boldsymbol{Y}, \underline{\boldsymbol{S}}\}$ conditioned on the value of $p_2 = \mathsf{P}\{X_2 = 1|\boldsymbol{Y}, \underline{\boldsymbol{S}}\}$. Without loss of genearality, the actual values of X_1 and X_2 were both equal to 1. For each given value of $\mathsf{P}\{X_2 = 1|\boldsymbol{Y}, \underline{\boldsymbol{S}}\}$, the curve can be regarded as the probability density function of $\mathsf{P}\{X_1 = 1|\boldsymbol{Y}, \underline{\boldsymbol{S}}\}$. The parameters are: the signal dimension $N = 500$, the number of measurements $M = 250$, the sparsity $\epsilon = 0.1$, the SNR per input element $\gamma = 10$ dB, and identical amplitude $U_n \equiv 1$ for all n. The sparse measurement matrix S is such that S_{mn} is -1 or $+1$, each with probability 0.02, and otherwise 0.

V. The Unambiguous Case

For relatively large SNR (we will discuss how large shortly), the MMSE can be approximated as follows. Consider a suboptimal estimator that first decides whether the variable X takes zero value and then estimates its value if X is believed to be nonzero. Suppose the error probability of the first decision is no greater than P_e regardless of whether X is zero or nonzero. The MMSE can be upper bounded:

$$\mathsf{mmse}(P_X, \Gamma) \le P_e + \epsilon \cdot \mathsf{mmse}(P_U, \Gamma) \tag{18}$$

for every $\Gamma \ge 0$, where the term P_e upper bounds the average error caused by mis-detection in the first step (recall that $\mathsf{E}\{U^2\} = 1$), whereas the second term approximates the error made when the variable is correctly detected to be nonzero.

Suppose the random variable U is lower bounded by d_{min} in its absolute value, i.e., $|U| \ge d_{min}$ with probability 1. We refer to this as the *unambiguous case*. It is not difficult to see that

$$P_e \le e^{-\gamma d_{min}^2/2}. \tag{19}$$

This implies that the fixed-point equation can be expressed as

$$\eta^{-1} = 1 + \frac{\gamma}{\mu}\mathsf{mmse}(P_U, \eta\gamma) + \frac{\gamma}{\epsilon\mu}e^{-\eta\gamma d_{min}^2/2}. \tag{20}$$

Note that

$$\frac{\gamma}{\mu}\mathsf{mmse}(P_U, \eta\gamma) \le \frac{\gamma/\mu}{1+\eta\gamma} \tag{21}$$

which is much smaller than 1 if $\gamma \ll \mu$ or $\eta\mu \gg 1$ (or both). Furthermore, for large enough $\eta\gamma$, the last term in (20) is also much smaller than 1. Under these circumstances, $\eta \approx 1$ by (20). It is clear that this is typically the case if one chooses $\mu \gg 1$. Moreover, in order for the approximation (18) to be accurate with $\Gamma = \eta\gamma$, one should also choose $\gamma \gg 1$, which is again easy to meet.

With enough measurements and SNR, the support set of the sparse signal can be determined with high fidelity. We have the following result as a special case.

Claim 4: Suppose for each $n = 1, \ldots, N$, U_n is equally likely to be ± 1, i.e., X_n takes the values ± 1 with probability $\epsilon/2$ and 0 otherwise. If $\min(\mu, \mu\gamma) \gg 1$, then the minimum probability of error for detection of X_n based on the measurements $(\boldsymbol{Y}, \underline{\boldsymbol{S}})$ is upper bounded by $\exp[-\mu\gamma/4]$, and the MMSE of estimating the sparse signal is upper bounded by $\epsilon\,\mathsf{mmse}(P_U, \eta\mu\gamma) + e^{-\eta\gamma/2}$ per dimension. If U is not binary but $|U| \ge d_{min}$ with probability 1, then the error probability is no larger than $\exp\left[-\mu\gamma d_{min}^2/4\right]$.

The bound is quite useful. Consider the following example: Let $N = 10,000$ and $\epsilon = 0.001$ so that the support set of the sparse signal is of cardinality 10 on average. Suppose the signal takes the value ± 1 if nonzero. Consider an SNR of 0 dB and 500 measurements in total (so that $\mu = 50$). The resulting error probability is no greater than $e^{-12.5} \approx 3.7 \times 10^{-6} < 1/N$. This suggests that one rarely makes any errors in terms of estimating the support set of the sparse signal. Furthermore, the MMSE for estimating $\boldsymbol{X}$ can be obtained as 8.6×10^{-6} per dimension, which amounts to 0.086 in total. Consider the alternative bound in [5], which is no better than

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2 \le 32\log(N)/(\mu\gamma) \approx 5.9\ . \tag{22}$$

This latter bound (22) does not prevent one from making 5 errors in terms of estimating the support set of the sparse signal, which is an error probability of 5×10^{-4}. It appears that the bound in [5] is pessimistic when applied to the Bayesian framework—by roughly two orders of magnitude in this case.

If P_U is an arbitrary distribution, which is not necessarily unambiguous, then the analysis can be more complex. One can still obtain useful upper bounds on the error probability if $|U|$ is bounded away from 0 with high probability. If the values of X_n flirt with 0

with non-negligible probability, then it is generally impossible to be accurate in estimating the support set of the sparse signal. Indeed, other authors (cf., Akçakaya and Tarokh [7] and references therein) discuss the case where the SNR must be increased if some signal values are near zero.

VI. Conclusion

This paper describes a fundamental single-letter characterization of the compressed sensing (CS) problem. Also discussed is a result on the decoupling of the elements of the sparse signal. Belief propagation is shown to often be asymptotically optimal in case of sparse measurements. If the replica method is justifiable, then using sparse measurement matrices performs as well as using dense measurement matrices. This suggests that for relatively large systems, one should prefer to use sparse measurement matrices so that low-complexity algorithms such as belief propagation can exploit the sparsity of the measurement matrix without sacrificing the estimation performance.

It is interesting to note that although this work considers independent measurement noise, the implications may apply to the case of quantization noise studied in [22], [23]. This is because, for a given SNR, additive independent Gaussian noise is often the worst case. The replica method is also applicable to suboptimal estimators [13], [15]. A possible direction of future work is to study the performance of various other CS algorithms in the literature, such as [24]–[27].

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[2] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[3] D. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9452–9457, 2005.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *arxiv:0907.3574v1*, 2009.

[5] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[6] S. Sarvotham, D. Baron, and R. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conf. Commun., Control, and Computing*, 2006.

[7] M. Akçakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," *arXiv:0711.0366v1*, 2007.

[8] W. Wang, M. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse support recovery: Dense versus sparse measurements," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 2197–2201, Toronto, Canada, 2008.

[9] P. Tune, S. R. Bhaskaran, and S. Hanly, "Number of measurements in sparse signal recovery," in *Proc. IEEE Int. Symp. Inform. Theory*, pp. 16–20, Seoul, Korea, 2009.

[10] S. Rangan, A. K. Fletcher, and V. K. Goyal, "Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing," *arXiv:0906.3234v1*, 2009.

[11] S. Aeron, M. Zhao, and V. Saligrama, "Information theoretic bounds to performance of compressed sensing and sensor networks," *arXiv:0804.3439v2*, 2009.

[12] S. F. Edwards and P. W. Anderson, "Theory of spin glasses," *Journal of Physics F: Metal Physics*, vol. 5, pp. 965–974, 1975.

[13] T. Tanaka, "A statistical mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2888–2910, Nov. 2002.

[14] D. Guo and S. Verdú, "Multiuser detection and statistical mechanics," in *Communications, Information and Network Security* (V. Bhargava, H. V. Poor, V. Tarokh, and S. Yoon, eds.), ch. 13, pp. 229–277, Kluwer Academic Publishers, 2002.

[15] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1982–2010, June 2005.

[16] D. Guo and T. Tanaka, "Generic multiuser detection and statistical physics," in *Advances in Multiuser Detection* (M. Honig, ed.), ch. 5, Wiley-IEEE Press, 2009.

[17] R. R. Müller, "Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation," *IEEE Trans. Signal Processing*, vol. 51, pp. 2821–2828, Nov. 2003.

[18] D. Guo and C.-C. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Select. Areas Commun.*, vol. 26, pp. 421–431, Apr. 2008.

[19] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, August 2001.

[20] D. Baron, S. Sarvotham, and R. G. Baraniuk, "Bayesian compressive sensing via belief propagation," *to appear in IEEE Trans. Signal Process.*, 2009.

[21] A. Montanari, "Estimating random variables from random sparse observations," *European Trans. Telecommun.*, vol. 19, pp. 385–403, Apr. 2008.

[22] E. Candes and J. Romberg, "Encoding the ℓ_p ball from limited measurements," in *Data Compression Conference*, pp. 33–42, 2006.

[23] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. Conf. Inform. Sciences & Systems*, Princeton, NJ, USA, 2008.

[24] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *arXiv:0803.0811v3*, 2009.

[25] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, pp. 301–321, 2009.

[26] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Top. Signal Process.*, vol. 1, pp. 586–597, 2007.

[27] P. Schniter, L. C. Potter, and J. Ziniel, "Fast Bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," *IEEE Trans. Signal Process.*, 2009.